MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 7: 37–49
DOI 10.3897/mbmg.7.98539

Research Article

# Maximizing the reliability and the number of species assignments in metabarcoding studies using a curated regional library and a public repository

**Audrey Bourret**[1], **Claude Nozères**[2], **Eric Parent**[1], **Geneviève J. Parent**[1]

1 *Laboratory of Genomics, Maurice Lamontagne Institute, Fisheries and Oceans Canada, 850 Rte de la Mer, Mont-Joli, QC, G5H 3Z4, Canada*

2 *Maurice Lamontagne Institute, Fisheries and Oceans Canada, 850 Rte de la Mer, Mont-Joli, QC, G5H 3Z4, Canada*

Corresponding authors: Audrey Bourret (audrey.bourret@dfo-mpo.gc.ca); Geneviève J. Parent (genevieve.parent@dfo-mpo.gc.ca)

**Academic editor:** Florian Leese | **Received** 7 December 2022 | **Accepted** 14 February 2023 | **Published** 23 February 2023

## Abstract

Biodiversity assessments relying on DNA have increased rapidly over the last decade. However, the reliability of taxonomic assignments in metabarcoding studies is variable and affected by the reference databases and the assignment methods used. Species level assignments are usually considered as reliable using regional libraries but unreliable using public repositories. In this study, we aimed to test this assumption for metazoan species detected in the Gulf of St. Lawrence in the Northwest Atlantic. We first created a regional library (GSL-rl) by data mining COI barcode sequences from BOLD, and included a reliability ranking system for species assignments. We then estimated 1) the accuracy and precision of the public repository NCBI-nt for species assignments using sequences from the regional library and 2) compared the detection and reliability of species assignments of a metabarcoding dataset using either NCBI-nt or the regional library and popular assignment methods. With NCBI-nt and sequences from the regional library, the BLAST-LCA (least common ancestor) method was the most precise method for species assignments, but the accuracy was higher with the BLAST-TopHit method (>80% over all taxa, between 70% and 90% amongst taxonomic groups). With the metabarcoding dataset, the reliability of species assignments was greater using GSL-rl compared to NCBI-nt. However, we also observed that the total number of reliable species assignments could be maximized using both GSL-rl and NCBI-nt with different optimized assignment methods. The use of a two-step approach for species assignments, i.e., using a regional library and a public repository, could improve the reliability and the number of detected species in metabarcoding studies.

## Key Words

classifier, cytochrome C oxidase I, GenBank, marine species, metagenomics, reference sequence library

## Introduction

Biodiversity assessments and monitoring using DNA have increased rapidly over the last decade given the high potential of this non-intrusive approach to uncover biodiversity efficiently with limited effort (Taberlet et al. 2012; Makiola et al. 2020). Environmental DNA (eDNA) metabarcoding surveys allow the detection of a diversity of organisms in various types of environmental samples using high-throughput sequencing (Taberlet et al. 2012; Yu et al. 2012). Surveys of eDNA generally involve a series of steps such as sample collection, extraction, targeted amplification, high-throughput sequencing, and bioinformatic processing, which includes taxonomic assignments to reference sequences from a public repository or a regional library (Deiner et al. 2017). Only a small fraction of detected eDNA sequences in environmental samples can currently be assigned to a species-level identity owing to a lack of data and taxonomic resolution in publicly available resources (Deiner et al. 2017; Leite et al. 2021; Zafeiropoulos et al. 2021). The reliability and precision of taxonomic assignments is affected by the quality and availability of sequences in repositories and the assignment methods, thereby limiting confidence in the use of eDNA

Copyright *Copyright work is made by His Majesty or by an officer or servant of the Crown in the course of his duties. This is an open access article distributed under the terms of the CC0 Public Domain Dedication.*

PENSOFT

for biodiversity monitoring and targeted species detections (Coissac et al. 2012; McGee et al. 2019; Meiklejohn et al. 2019; Gold et al. 2021; Hleap et al. 2021).

Several public repositories exist and can be used as reference databases to provide taxonomic assignments in metabarcoding studies. The public National Center for Biotechnology Information Nucleotide database (NCBI-nt, including the GenBank database) is the largest sequence repository and is widely used in eDNA metabarcoding studies (Porter and Hajibabaei 2018b, 2020). However, the presence of mislabeled specimens, the large variation in quality of sequences available, and gaps in species coverage (i.e., unrepresented species) result in erroneous species identification when directly comparing unknown sequences to NBCI-nt (Bidartondo 2008; Mioduchowska et al. 2018; Leray et al. 2019). The Barcode of Life Data Systems (BOLD) is another sequence repository specific to the most common barcode regions, including the cytochrome c oxidase I (COI) gene, which is the widely used gene region for animal DNA barcoding (Ratnasingham and Hebert 2007; Porter and Hajibabaei 2018b). BOLD displays mandatory (e.g., institution storing voucher specimen, sampling country) and optional (e.g., sampling location, specimen photos) metadata, performs groupings of similar sequences into Barcode Index Numbers (BINs), and permits editing or updating of records, all of which assists with data quality control. However, like NCBI-nt, it is also vulnerable to submissions of misidentified specimens (McCusker et al. 2013; Oliveira et al. 2016; Fontes et al. 2021; Radulovici et al. 2021). As reliable taxonomic assignments at the species-level are expected under many regulatory contexts (e.g., environmental status assessment, monitoring of invasive species or species at risk; Aylagas et al. 2014; Hering et al. 2018; Bush et al. 2019; Piper et al. 2019), some metabarcoding studies have questioned the value of using public repositories (e.g., von Ammon et al. 2018; Locatelli et al. 2020; Gold et al. 2021). Nonetheless, these important resources remained essential for identification of sequences of unknown origin, and could be valuable starting points for the creation of smaller curated reference sequence libraries. Characterizing the proportion of accurate species assignments using NCBI-nt would be highly valuable to understand the extent of uncertainty in metazoan species' eDNA detection and consequently enable an accurate interpretation of metabarcoding results.

Alternatively, curated regional libraries have been shown to reduce errors in species assignments (Gold et al. 2021). Regional libraries are limited to species expected in predefined areas, and can be created by data mining and curating existing sequences from public repositories and/or from generating sequences from specimens. They have the advantage of limiting spurious assignments to related but non-local species and to reveal gaps (i.e., missing sequences) in taxonomic groups (Weigand et al. 2019; Ramirez et al. 2020; Jazdzewska et al. 2021). Examples of regional libraries are available in the northern hemisphere for multiple taxonomic groups (e.g., freshwater fish: Knebelsberger et al. 2014; Hänfling et al. 2016; marine fish: Stoeckle et al. 2017; Fraija-Fernández et al. 2020; Gold et al. 2021; macrobenthos: Van Den Bulcke et al. 2021). Some of these reference libraries present ranking systems to ensure high taxonomic reliability (e.g., from Grade A for highest reliability to E for lowest reliability, Costa et al. 2012; see Knebelsberger et al. 2014 for an example of its application on fish sequences). Ranking systems are often provided to target future barcoding efforts and improvements in reference sequences. No explicit ranking system about the uncertainty of species assignment has yet been presented within metabarcoding studies. Such a system would be highly valuable to provide clear indications on the reliability of species assignments for eDNA end-users.

Another source of variability in species' assignments is the bioinformatics software and pipelines used in metabarcoding studies. Recently, studies have started to evaluate the accuracy of taxonomic assignments using various bioinformatic methods (O'Rourke et al. 2020; Hleap et al. 2021; Mathon et al. 2021). These studies compared taxonomic assignment methods that are based on different strategies, such as alignment, composition, or modelling (Richardson et al. 2017, see also four strategies in Hleap et al. 2021). The Basic Local Alignment Search Tool (BLAST) is an alignment-based approach extensively used in metabarcoding studies that relies on similarity between unknown sequences and records from a reference database to return best hits (Camacho et al. 2009). The taxonomic identity of the unknown sequence may be inferred in conjunction with a least common ancestor (LCA) or a Top Hit approach with identity threshold, usually between 95 and 99%. These thresholds should reflect expected inter-species divergence, but high variation among taxonomic groups may cause pitfalls in assignments (Wang et al. 2007; Alberdi et al. 2018). Composition-based classifiers that involve machine-learning algorithms such as Ribosomal Database Project (RDP; Wang et al. 2007) and IDtaxa (Murali et al. 2018) have shown good performance for species' assignments (Richardson et al. 2017; Murali et al. 2018; Porter and Hajibabaei 2018a). They use the frequency or weighted frequency of k-mers (i.e., short unique sequence substrings) to compare the composition of a query sequence to reference sequences, and then provide a measure of confidence for a taxonomic assignment through bootstrapping the assignment process. They are thus less affected by low divergence between groups. Supervised classifiers are trained on a reference library, and pre-trained classifiers are increasingly available (e.g., Porter and Hajibabaei 2018a). However, recent benchmarking studies have shown lower performance of such classifiers compared to BLAST (O'Rourke et al. 2020; Hleap et al. 2021; Mathon et al. 2021).

This study aimed to estimate the accuracy and precision of species assignments using the public repository NCBI-nt and to contrast the reliability of using NCBI-nt and a regional library for species assignments of a metabarcoding dataset, with popular taxonomic assignment methods (Fig. 1). Specifically, we first created a curated regional library (GSL-rl) using publicly available sequences from BOLD for the COI barcode locus of metazoans

from the Gulf of St. Lawrence (Fig. 1A). The regional library also contained a reliability ranking system for species assignments based on sequence availability and similarity that can be understood by any eDNA end-users, regardless of their scientific expertise. Then, we used the reference sequences from the GSL-rl to estimate the accuracy and precision of NCBI-nt (Fig. 1B). We also compared the detected species in a metabarcoding dataset when using NCBI-nt or GSL-rl, and we contrasted their reliability (Fig. 1C). These two sets of reference sequences likely represented two extreme scenarios in terms of curation and size. We reached the conclusion that using a two-step approach, i.e., species assignments first with a regional library and then with a public repository to contrast results, is desirable to maximize the reliability and number of species assignments in metabarcoding studies.

# Methods

### Creation of a regional library for the Gulf of St. Lawrence (GSL-rl) with a reliability ranking system

The creation of a curated regional library for the Gulf of St. Lawrence (hereafter GSL regional library: GSL-rl) was done through multiple rounds of data mining on BOLD for marine metazoan species (i.e., vertebrate and invertebrate) and revisions based on quality and similarity of sequences (Fig. 1A, see also Suppl. material 1: fig. S1 for more details). The initial list of species was obtained from 1) decision-makers, and 2) regional species' list with taxonomical information (Nozères 2017), and was extended along the creation process. All sequences retained in the GSL-rl had names at the genus or species levels and were already published on BOLD, and were not from outgroup taxa (e.g., human, plant, bacteria) nor hybrid species. We computed a genetic distance intra and interspecific between BINs as the Kimura's 2-parameters distance (Kimura 1980) with ape R package (v.5.0, Paradis and Schliep 2019). More details about the creation of the GSL-rl are provided in the Suppl. material 1 (see also Suppl. material 1: tables S1, S2).

We created the GSL-rl to identify molecular operational taxonomic units (MOTUs) at the species level. Each species in the GSL-rl was ranked based on sequence availability and similarity (Fig. 1A, Suppl. material 1: table S3). Species with reference sequences for itself and closely related species (i.e., from the same genus) acknowledged to be present in the Gulf of St. Lawrence were ranked as "Reliable" if they did not share BOLD's barcode index number (BIN; i.e., a unique identifier of sequences based on genetic distance, Ratnasingham and Hebert 2013). Species with reference sequences for themselves, but not for all congeners acknowledged to be present in the Gulf of St. Lawrence (i.e., other species of the same genus), were ranked as "Unreliable due to gaps". Species with reference sequences sharing BIN with other species were ranked as "Unreliable due to BIN sharing". Common causes of BIN sharing are genetic similarities between species or specimen misidentification. For the GSL-rl, the curation and validation process done during its creation should limit the BIN sharing due to specimen misidentification. Taxonomic assignments belonging to one of the two "Unreliable" categories should be interpreted with caution and preferably not at the species-level.

### Evaluating the accuracy and precision of species assignments using the public NCBI-nt repository

We used the curated sequences from the GSL-rl to evaluate the accuracy and precision of species assignments using NCBI-nt (Fig. 1B). Taxonomic assignments were performed over the sequences contained in GSL-rl using NCBI-nt (downloaded 2020-10-23) and the BLAST+ tool blastn (v2.10.1, Camacho et al. 2009) combined with the least common ancestor (LCA; hereafter BLAST-LCA) or the Top Hit methods (hereafter, BLAST-TopHit) at three identity thresholds (95, 97 and 99%) with an in-house R script. The LCA method assigns the higher taxonomic rank shared by all hits above the identity threshold. Top Hit method assigns the higher taxonomic rank shared by the hits with the lowest e-value. We excluded hits containing "environmental sample", "uncultured" or "predicted" in their description.

We evaluated two performance parameters, i.e., accuracy and precision, for species assignments using NCBI-nt. To compute these parameters, we classified each taxonomic assignment (following Bokulich et al. 2018) as either:

- A true positive (TP), or accurate species assignment, if the assignment was with the correct taxonomical classification, e.g., an *Ammodytes hexapterus* sequence correctly identified as is.
- A false positive (FP), or inaccurate species assignment, if the assignment was with an incorrect taxonomical classification, e.g., an *Ammodytes hexapterus* sequence incorrectly identified as *Ammodytes marinus*.
- A false negative (FN) if the assignment was at a taxonomical level higher than species, no matter if the assignment was correct or not, e.g., an *Ammodytes hexapterus* sequence classified as *Ammodytes* sp. This is equivalent to an under-classification error (Edgar 2018).

The accuracy, reflecting the proportion of accurate assignments at the species level, was defined as TP / (TP + FP + FN), whereas the precision was defined as TP / (TP + FP).

### Contrasting species assignments using the regional library or the public NBCI-nt repository, and popular assignment methods.

We compared the detection results from an eDNA metabarcoding dataset using GSL-rl and NCBI-nt and three assignment methods (Fig. 1C). The eDNA metabarcoding dataset was obtained from the analysis of 2L water sam-

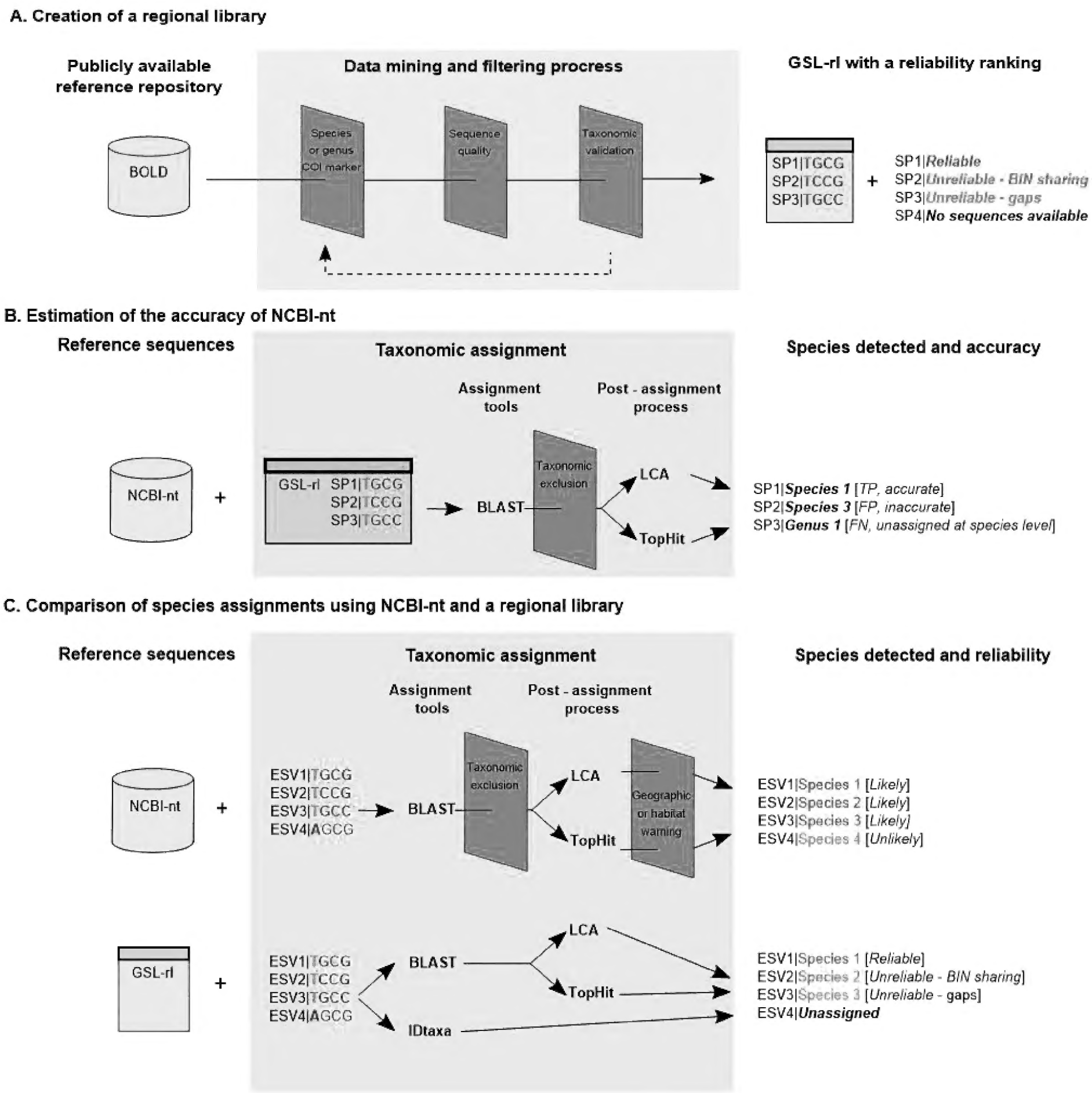


**Figure 1.** Schematic representation of the three main steps of this study. **A** Creation of a regional library for metazoans from the Gulf of St. Lawrence (GSL-rl). Sequences were selected from BOLD and curated through multiple filtering and auditing steps (see Fig. S1 for more details on filtering and auditing parameters). Species or genus were added through an iterative process to further improve the regional library. Each species in the GSL-rl was ranked based on sequence availability and sequence similarity to closely related species in the Gulf of St. Lawrence; **B** Estimation of the accuracy and precision of NCBI nucleotide database (NCBI-nt) using the reference sequences from the regional library. Taxonomic assignments were performed using NCBI-nt over the reference sequences from GSL-rl, using the Blast+ tool *blastn* (hereafter BLAST; Camacho et al. 2009). Assignment results were filtered based on taxonomic identity, then a least common ancestor (LCA) or a TopHit method were used to assign a unique taxon identity to each sequence. Each assignment was then classified as a true positive (TP, accurate), a false positive (FP, inaccurate) or a false negative (FN, unassigned at the species level). Performance parameters were derived from this classification; **C** Comparison of species assignments and their reliability using NCBI-nt or GSL-rl. Taxonomic assignments of ESVs from a metabarcoding dataset were performed with BLAST and with the classifier IDtaxa (Murali et al. 2018). For NCBI-nt, the species ranking involved a plausibility filter based on the location. For GSL-rl, the species ranking was directly provided with the library (see methods for more details).

ples (n=61) collected from scientific surveys in 2018 in coastal areas of the Gulf of St. Lawrence (Suppl. material 1: fig. S2), both at the surface and bottom of the water column. Water samples were filtered on glass fiber filter (47 mm, 1.2 and/or 10 µm pore size; Sigma Aldrich, MO, U.S.) in an ultraclean room, and DNA was extracted from filters using Qiagen Blood and Tissus Kit (QIAGEN, MD, U.S.). Genome Québec performed PCR amplifications with the primers mICOIintF (Leray et al. 2013) and jgHCO2198 (Geller et al. 2013), targeting a 313 pb section

of the COI Folmer region. Each amplification was then indexed and pooled. Amplicons pools were sequenced using Illumina MiSeq PE 250 bp at Genome Québec. Negative controls added at each step (n = 22) were also sequenced. The bioinformatic pipeline involved the use of dada2 R package (Callahan et al. 2016) and a correction of the ESV table based on negative samples (Suppl. material 1: fig. S3). See supplementary material for more details on the field, laboratory and bioinformatics works underlying the eDNA metabarcoding dataset.

The three assignment methods compared were BLAST-LCA, BLAST-TopHit and IDtaxa. BLAST assignment methods were used as described in the previous section with both GSL-rl and NCBI-nt. NCBI-nt BLAST results were filtered to retain only metazoan detections and remove non-marine taxa (i.e., *Homo sapiens*, Arachnida, Insecta). IDtaxa is a classifier implemented within the DECIPHER R package (Wright 2016) and was trained only with the GSL-rl. The training of IDtaxa directly over the full NCBI-nt would have been too computationally intensive, and would have required a minimal curation to restrict the scope to metazoan sequences (see also Porter and Hajibabaei 2018a for a curated version of public repositories' training set). The IDtaxa classifier was selected since it has been shown to be less prone to "over-classification", i.e., classification to an erroneous group when the real group is absent from the training set, compared to the popular RDP classifier (Murali et al. 2018). Taxonomic assignments with IDtaxa were obtained at three confidence thresholds (i.e., weighted fraction of bootstrap replicates assigned to a given taxa) representing moderate confidence (40%), high confidence (50%), and very high confidence (60%) in species assignments (Murali et al. 2018).

We contrasted results obtained using GSL-rl and NCBI-nt with distinct ranking systems (Fig. 1C). Species detected with the GSL-rl were classified according to the three categories of the reliability ranking system previously created: "Reliable", "Unreliable due to gaps", "Unreliable due to BIN sharing" (Suppl. material 1: table S3). For species assignments with NCBI-nt, we used geographic and habitat filters to classify them as "Likely" if they were part of the Gulf of St. Lawrence checklist (Nozères 2017) or present in the areas based on the World Register of Marine Species (WoRMS, WoRMS Editorial Board 2020), and "Unlikely" if not. Such filters are often applied in metabarcoding studies but the source of information for the likelihood of a species to be present is often obscure.

## Data accessibility

Raw sequence data from the metabarcoding dataset are available in the Sequence Read Archive (SRA) under the accession number PRJNA925571.

The data and scripts used in this manuscript are stored in the github repository https://github.com/GenomicsMLI-DFO/GSL_COI_ref_library The GSL-rl database (sequences, reliability ranking and trained dataset) can be found in the github repository https://github.com/GenomicsMLI-DFO/MLI_GSL-rl.

# Results

### A COI regional library with a reliability ranking system for metazoans from the Gulf of St. Lawrence (GSL-rl)

The first version of GSL-rl comprised 1304 sequences covering 439 species (158 species from the phylum Chordata spanning 68 families; 281 species of invertebrates spanning 129 families and 9 phyla) and 11 other taxa at the genus level only (Vertebrates: 3 genera from 2 families; Invertebrates: 8 genera from 8 families and 4 phyla; Fig. 2). It represented 67.4% of the taxa on the target list (651 species; Vertebrates: 94.6%; Invertebrates: 58.1%; Suppl. material 1: table S3). The sequences were retrieved mostly from the Northwest Atlantic Ocean (67.8%). A total of 525 BINs were represented (Vertebrates: 159; Invertebrates: 366), with 16 BINs that were shared by at least two taxa (Vertebrates: 8; Invertebrates: 8; Suppl. material 1: table S4), and 58 taxa occupied more than one BIN (6 vertebrates with up to 3 BINs; 52 invertebrates with up to 7 BINs; Suppl. material 1: table S5). The median sequence length was 658 bp (range: 640–664 pb) while the mean (± SD) of missing values (N's) was 0.002 ± 0.034% (max < 1%). Genetic distances were on average 0.005 (range: 0.000–0.023) within BINs and 0.122 (range: 0.012–0.347) between intraspecific BINs.

We then provided a reliability ranking for each species within GSL-rl based on the completeness of sequences available (Fig. 2, Suppl. material 1: table S6; see methods for more details). Species within the "Reliable" category accounted for the largest proportion of the species with sequences from the regional library (302 species or 68.8%; 133 vertebrates, 169 invertebrates). Species classified to the "Unreliable due to BIN sharing" and the "Unreliable due to gaps" categories represented 5.2% (23 species; 13 Chordata, 10 invertebrates) and 26.0% (114 species; 12 vertebrates, 102 invertebrates) of the GSL-rl species, respectively. The GSL-rl database (version 1.0 and future versions) is available on GitHub (https://github.com/GenomicsMLI-DFO/MLI_GSL-rl).

### Accuracy and precision of species assignments using NCBI-nt and two assignment methods

The proportions of species assignments over all taxa were higher with the BLAST-TopHit method (range: 85.5–87.9%) than the BLAST-LCA method (range: 47.6–71.0%) with any identity thresholds (Fig. 3A). Over all taxa, the proportions of species assignments increased with the BLAST-LCA method while they decreased with the BLAST-TopHit method with increasing identity thresholds (Fig. 3A). Across taxonomic groups, the proportions of species assignments were also consistently greater at all thresholds for the Blast-TopHit method compared to the BLAST-LCA method. The proportions of species assignments at the 97% similarity threshold varied with the BLAST-TopHit method from 74.4% for

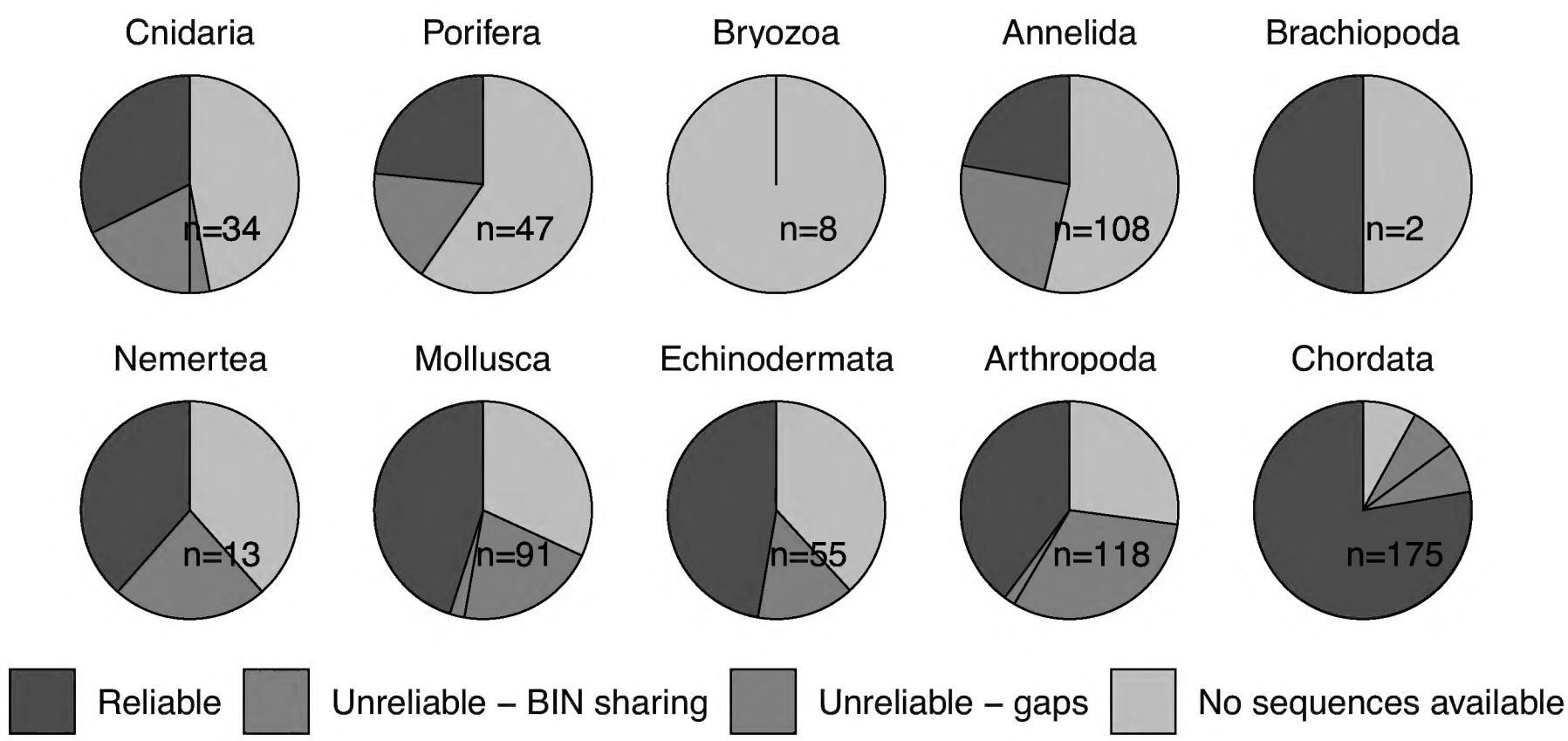


**Figure 2.** Classification of 651 marine metazoan species previously observed in the Gulf of St. Lawrence and included in GSL-rl, by phylum. Species reliability ranking is based on the availability from local species and sequence similarity to closely related species in the Gulf of St. Lawrence.

Annelida, Brachiopoda, and Nemertea to 93.1% for Arthropoda method and with the BLAST-LCA method from 35.8% for Cnidaria and Porifera to 75.2% for Arthropoda (Fig. 3B).

The accuracy, or proportion of accurate species assignments, was higher with the BLAST-TopHit method compared to the BLAST-LCA method, over all taxa and in each taxonomic group at all identity thresholds (Fig. 3A, B). Over all taxa, accuracy varied between 80.1% and 82.5% for the BLAST-TopHit method and between 42.7% and 68.0% for the BLAST-LCA method for the three identity thresholds tested (Fig. 3A). For each taxonomic group, the accuracy was consistently higher at all thresholds with the BLAST-TopHit method compared to the BLAST-LCA method. The accuracy at the 97% threshold varied with the BLAST-TopHit method from 69.6% for Annelida, Brachiopoda and Nemertea to 89.6% for Arthropoda and with the BLAST-LCA method from 34.0% for Cnidaria and Porifera to 73.6% for Arthropoda (Fig. 3B).

The precision was greater for the BLAST-LCA method compared to the BLAST-TopHit method over all taxa at all thresholds (BLAST-LCA range: 95.7–96.9%, BLAST-TopHit range: 93.8–94.4%; Fig. 3A) and in most taxonomic groups at the 97% threshold (BLAST-LCA range: 92.3–99.2%, BLAST-TopHit range: 89.6–96.3%; Fig. 3B).

## Comparison of the reliability of species assignments to a metabarcoding dataset using GSL-rl and NCBI-nt with three assignment methods

We used an eDNA metabarcoding dataset to compare the number and the reliability of species assigned using GSL-rl and NCBI-nt with three assignment methods. The five possible combinations of repository/library and assignment methods were GSL-rl and NCBI-nt with BLAST-LCA (1, 2), GSL-rl and NCBI-nt with BLAST-TopHit (3,4), and GSL-rl with IDtaxa (5; Fig. 1C). A total of 80 species were assigned with the five combinations of repository/library and assignment methods (Fig. 4A). Detected species differed using NCBI-nt and GSL-rl and the three assignment methods (Fig. 4A).

Across all combinations, the highest and lowest numbers of species assigned were observed with NCBI-nt and BLAST-TopHit95 (66 species) and BLAST-LCA95 (44 species), respectively (Fig. 4B). The number of assigned species decreased with increasing thresholds for most combinations, except for BLAST-LCA with NCBI-nt (Fig. 4B). For GSL-rl, proportions of assigned species ranked as "Unreliable due to BIN sharing" or "Unreliable due to gaps" did not change directionally with increasing thresholds (Fig. 4B). For NCBI-nt, decreasing proportions of "Unlikely" species were assigned with increasing identity thresholds of BLAST-LCA or BLAST-TopHit (Fig. 4B).

The assignment method with the maximum number of assigned species differed between GSL-rl and NCBI-nt. The maximum number of assigned species was 62 species with GSL-rl and IDtaxa40 and 66 species with NCBI-nt and TopHit95 (Fig. 4B). Out of the 62 species assigned using the GSL-rl/IDtaxa40 combination, 46 species (74.2%) were ranked as "Reliable". The remaining assigned species were ranked as "Unreliable due to BIN sharing" (4 species, 6.5%) or "Unreliable due to gaps" (12 species, 19.4%; Fig. 4B). With the NCBI-nt/TopHit95 combination, 58 (87.9%) and 8 (12.1%) assigned species were ranked as "Likely" and "Unlikely" present, respectively (Fig. 4B).

Large proportions of detected species were exclusively assigned using only GSL-rl or NCBI-nt. A total of 30 species (37.5% of all species detected) were assigned

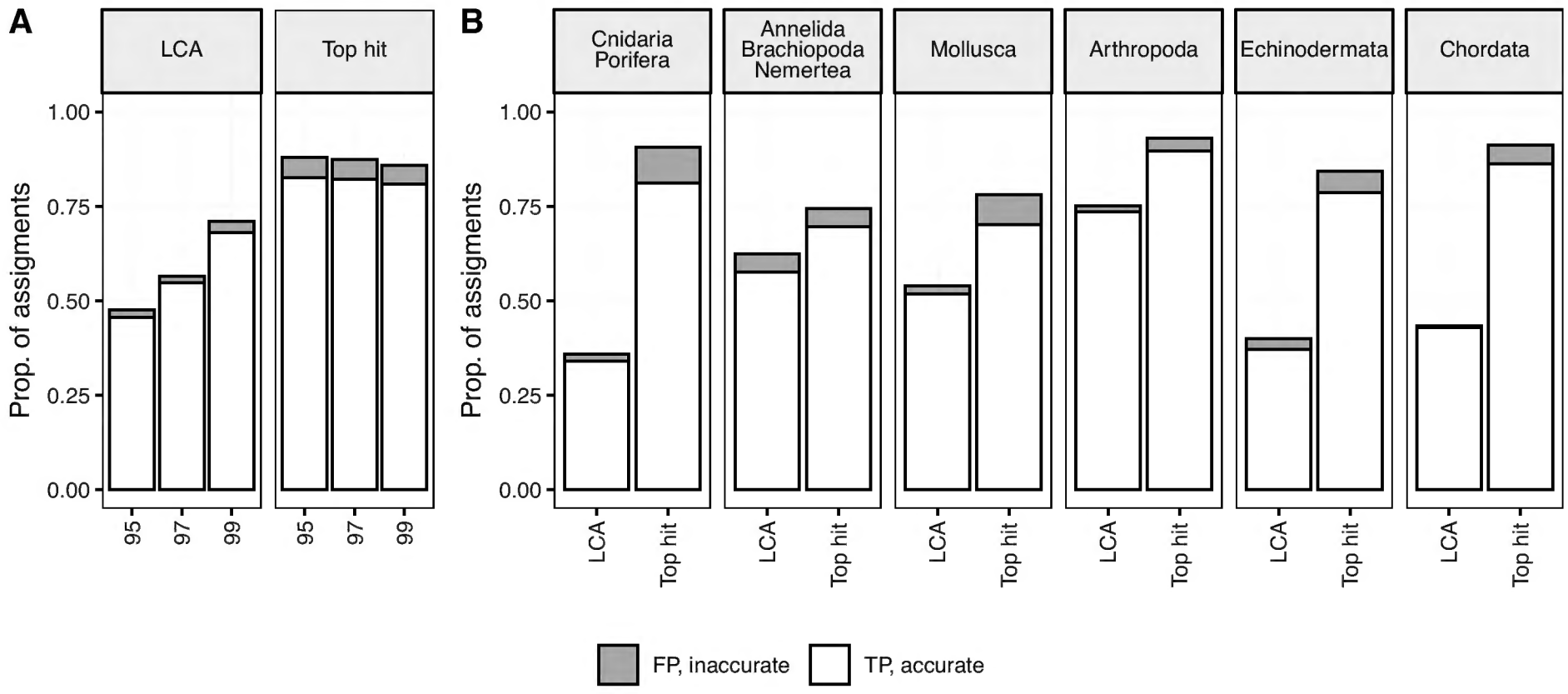


**Figure 3.** Taxonomic assignment results of sequences from GSL-rl using NBCI-nt and the BLAST-LCA and BLAST-TopHit methods. Proportions of accurate (true positive, TP) and inaccurate (false positive, FP) species assignments are presented **A** for all taxonomic groups at the three identity thresholds (95%, 97%, 99%) and **B** by taxonomic group at the 97% threshold.

only using GSL-rl (12 species) or NCBI-nt (18 species; Fig. 4A, C). For the species only assigned with GSL-rl, 7 species were ranked as "Reliable" whereas 1 and 4 species were ranked as "Unreliable due to BIN sharing" and "Unreliable due to gaps", respectively (Fig. 4A, C). For the species only assigned with NCBI-nt, 10 species were considered likely to be present in the GSL whereas 8 species were considered unlikely to be present. For the other 50 species assigned with both GSL-rl and NCBI-nt, 39 species were ranked as "Reliable" with the GSL-rl (78.0%, Fig. 4C). The remaining species assigned belonged to the "Unreliable due to BIN sharing" (3 species, 6.0%) or the "Unreliable due to gaps" categories (8 species, 16.0%; Fig. 4A, C).

# Discussion

### A COI regional library with a reliability ranking system for metazoans from the Gulf of St. Lawrence (GSL-rl)

The GSL-rl provides explicit reliability rankings for 651 species observed within the Gulf of St. Lawrence. We used two simple, broad categories, "Reliable" and "Unreliable", to characterize the robustness of species assignments in eDNA metabarcoding studies. The "Reliable" category represented the vast majority of species with reference sequences (68.8%, 302 species) in GSL-rl. Similar results were obtained for marine fish species from Portugal with the COI locus (73.5%, grade A, Costa et al. 2012). Past studies have shown the importance of a ranking system to limit erroneous species assignments (e.g., Costa et al. 2012; Knebelsberger et al. 2014). However, the ranking systems used in these studies are targeting an audience of barcoding specialists. With the mainly ranking system of species assignments in GSL-rl, we aimed to keep this classification simple to reach the large audience of eDNA users. Still, we used two "Unreliable" subcategories to highlight 1) the taxa necessitating future barcoding efforts, and 2) the relevance of the COI barcode to discriminate species. This allows any eDNA scientist to consider alternative loci if species of interest are not discriminated by the COI locus. Note that the reliability ranking of species in GSL-rl may change over time, particularly for understudied species. In the future, species may be upgraded to the "Reliable" category when further sequencing results fill in data gaps. Some species may also be downgraded to the "Unreliable" category, particularly for complex taxonomic groups in the region that should be targeted for review (e.g., polychaete worms).

The GSL-rl contains reference sequences for 439 species of the 651 targeted species of interest for conservation in the Gulf of St. Lawrence (i.e., 67.4%), with reference sequences available for a relatively large proportion of invertebrates (i.e., 59.1%). In Europe, marine invertebrates represent the taxonomic group with the lowest barcode coverage, and only 22.1% have one or more sequences available (Weigand et al. 2019). The larger proportion of invertebrates with reference sequences in GSL-rl is likely due to the species selection to initiate this regional library but also to the smaller study area and the barcoding campaigns for invertebrates in the Northwest Atlantic (e.g., Radulovici et al. 2009; Layton et al. 2016). Still, the GSL-rl is in its early development (v.1.0) and presently covers only a quarter of the estimated 2200 marine metazoan species that may occur in the Gulf of St. Lawrence (Nozères 2017). A review every two years is planned to increase the number of species covered by the GSL-rl.

The GSL-rl-could also improve species assignments in eDNA metabarcoding studies of the Northwest Atlantic and the Arctic Oceans compared to large public databases. The Gulf of St. Lawrence is a transitional

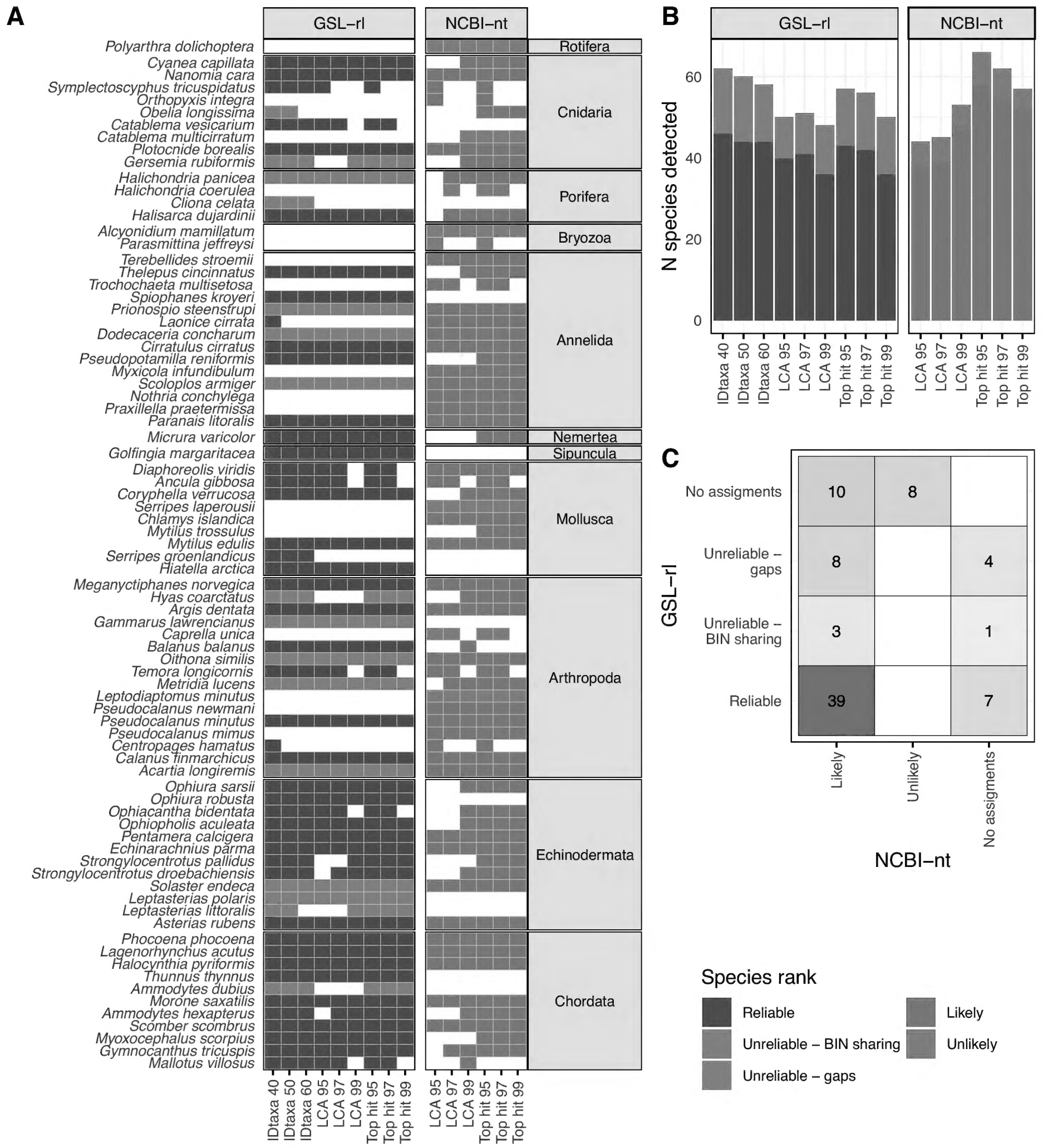


**Figure 4.** Assignment results at the species level using a regional library (GSL-rl) or a public repository (NBCI-nt) and popular assignment methods. We used three assignment methods, namely IDtaxa (confidence levels: 40%, 50% and 60%), BLAST-LCA, and BLAST-TopHit (identity thresholds: 95%, 97% and 99%). **A** Detections for each species and **B** number of species assignments for each source of reference sequences and method; **C** Comparison of species rank for all the species assigned with the two sources. Species rank categories are based on sequence availability and sequence similarity to closely related species in the Gulf of St. Lawrence for GSL-rl and on the geographic plausibility for NCBI-nt.

marine region where temperate southern species may occur alongside boreal and arctic species (Bourdages et al. 2022). There are no regional libraries covering marine metazoan species at the COI locus in nearby regions, and GSL-rl could promote the creation of these regional libraries. These can be created by data mining and curating existing sequences from public repositories (e.g., the approach used in this study), completely de novo from barcoding local specimens (e.g., Delrieu-Trottin et al. 2019), or from a combination of both approaches (e.g., Stoeckle et al. 2020; Gold et al. 2021). New tools are now emerging to facilitate the creation of regional reference libraries (e.g., Meta-Fish-lib, Collins et al. 2021; Barcode, Audit & Grade System (BAGS), Fontes et al. 2021; see also Kaehler et al. 2019 for a tool incorporating species abundance information into species assignments). Some

of the tools, such as BAGS, even allow for the annotation of species based on concordance between morphological species-based identification and sequence clusters in BOLD (Fontes et al. 2021). Combined with tools to find gaps in reference sequence libraries (e.g., GAPeDNA, Marques et al. 2021), more comprehensive species-level assignments are now possible.

## Accuracy of species assignments using NCBI-nt and two assignment methods

We estimated two performance parameters for metazoan species assignment using NCBI-nt, and observed large variations in results of performance parameters with the two assignment methods tested. While the BLAST-LCA method provided overall higher precision in species assignments, the accuracy was greater with BLAST-TopHit, an observation in line with a previous study (Hleap et al. 2021). The sensitivity of both methods to the prevalence of BIN sharing, gaps, and mislabeling within public repositories may explain the difference between performance parameters' results. For instance, the BLAST-LCA method, which is more conservative, is expected to generate more false negatives causing under-classification (i.e., assignments at a higher taxonomic level) in the presence of closely related species and BIN sharing. In contrast, the BLAST-TopHit method favors more species level assignment, but is highly impacted by gaps and mislabeled sequences (e.g., Schenekar et al. 2020), generating more false positives and lowering the precision of species assignments. We also want to highlight that the prevalence of gaps for targeted species may have been considerably reduced using sequences from the GSL-rl. The latter were retrieved from BOLD, which shares many records with NCBI-nt (Porter and Hajibabaei 2018b). Lower performances could be expected using a metabarcoding dataset, particularly for the TopHit method more sensitive to gaps. Nonetheless, the relatively good performance of the BLAST-TopHit method observed here suggests that misidentified specimens within NCBI-nt are limited for the targeted metazoan species of the Gulf of St. Lawrence.

Previous studies have shown that assignment methods can affect taxa detected in metabarcoding studies (O'Rourke et al. 2020; Hleap et al. 2021). As expected, the BLAST-TopHit method outperformed the BLAST-LCA method with NCBI-nt to provide a higher proportion of assignments (see also Hleap et al. 2021). Our results also showed that the proportion of accurate species assignments varied largely between taxonomic groups. Relatively well-described marine taxonomic groups such as Arthropoda (i.e., crustaceans) and Chordata (i.e., fishes and mammals) have reached accuracy ≥85% with the BLAST-TopHit method. The accuracy is much lower for the BLAST-LCA approach within the Chordata (43%), probably because this approach is more sensitive to the presence of close relative species (i.e., BIN sharing), reducing the potential of species level identification. Other groups, such as Annelida, Brachiopoda, Nemertea, and Mollusca, achieved lower accuracy using both the BLAST-TopHit and BLAST-LCA methods (maximum 70%).

The accuracy and precision using the sequences from GSL-rl in our study will be different at the time of reading this article due to the continuous growth of the public repository NCBI-nt. The publication of new sequences of low quality or with incorrect species identification can create unexpected ambiguities in species assignments as public repositories grow (Locatelli et al. 2020; Radulovici et al. 2021). Without a comprehensive versioning system, changes in the NCBI-nt database also limit the reproducibility of species assignments as it is difficult to identify and access a specific daily release. Note that starting with BLAST v.2.13 launched in March 2022, it is now possible to generate a metadata file describing the database used (Camacho and Madden 2022), which is an important step toward higher traceability.

## Comparing the reliability of species assignments to a metabarcoding dataset using GSL-rl and NCBI-nt with three assignment methods

The method with the maximum number of species assigned to the metabarcoding dataset differed between GSL-rl and NCBI-nt. The IDtaxa40 assignment method provided the highest number of species assigned using GSL-rl. Sequence composition strategies for species assignments, such as IDtaxa and RDP, had contrasting performance results in previous benchmarking studies (O'Rourke et al. 2020; Hleap et al. 2021; Mathon et al. 2021). Our results contrast with those from a previous study showing that IDtaxa did not perform as well as BLAST with mock communities composed of various freshwater taxonomic groups (Hleap et al. 2021). The contrasting results between the latter and our study could be explained by the difference in the confidence threshold used (Hleap et al. 2021). Parameter tuning may be key to choosing an optimal method for a dataset while more benchmarking studies are undertaken to develop parameter standards. The relatively better performance of IDtaxa in our study might also be due to the quality of the regional library used to train the classifier. Little is known about the impact of using classification training sets with varying levels of curation for taxonomic assignment, and the possible improvement in classifications when using regional libraries might be important in this context. With NCBI-nt, the number of detected species was greater with the BLAST-TopHit method compared to the BLAST-LCA method for the metabarcoding dataset. These results are similar to those obtained with GSL-rl COI sequences and have been discussed in the previous section.

More than a third of the species assigned to the metabarcoding dataset (n = 33 out of 80) were exclusive to either GSL-rl or NCBI-nt . For GSL-rl, the exclusion of non-indigenous species or mislabeled sequences increased the number of species assigned, confirming previous studies' results improving species assignments with regional libraries (von Ammon et al. 2018; Gold et al. 2021). The exclusion of non-indigenous species

increased the taxonomic resolution of the Atlantic bluefin tuna *Thunnus thynnus* within GSL-rl. Under-classification is usually observed when using NCBI-nt, as the Atlantic bluefin tuna presently shares a BIN (BOLD: AAA7352) with other *Thunnus* species that are not expected to be present in the Gulf of St. Lawrence (Nozères 2017). With NCBI-nt, detections in the "Likely" category comprised species for which sequences were not included in GSL-rl because of the stringency of quality filtering performed (e.g., Iceland scallop *Chlamys islandica*). Other detections in the "Likely" category were included in GSL-rl (e.g., the polychaete worm *Terebellides stroemii*), but the inability to detect them suggests that their intra-specific diversity is not fully covered by GSL-rl. Finally, a few species assigned with NCBI-nt were not listed as present in the Gulf of St. Lawrence, but after reconsideration, we concluded that they are likely to be found in the target area (e.g., *Pseudocalanus newmani*). All undetected species will be reviewed prior to the next release of GSL-rl. With NCBI-nt, we also observed under-classification of the sea star genus *Leptasterias* due to sequence mislabeling, which has been shown previously (Bidartondo 2008; Mioduchowska et al. 2018). The under-classification is due to two misidentified sequences, one is for *Leptasterias littoralis* identified as the sea star *Asterias forbesi* and the other is for *Leptasterias polaris* identified as the butterfly *Polyommatus fulgens*.

Comparing the ranking categories of NCBI-nt and GSL-rl revealed an important improvement in reliability with our annotated regional library (Fig. 1C). With NCBI-nt, we provided the likeliness of a species to be present in the Gulf of St. Lawrence due to the availability of a public species list (Nozères 2017). Such information is often difficult to obtain without expert knowledge (Pappalardo et al. 2021). Of all the species ranked in the "Likely" category using NCBI-nt, around 78% were classified as "Reliable" in the GSL-rl. Our results showed that the remaining 22% should be interpreted with caution given gaps (16%) or BIN sharing with close relative species (6%; Fig. 4C). Our results suggest that species level assignments of a metabarcoding dataset using NCBI-nt and a filter based on geographic plausibility can be misleading. This important hidden and overlooked uncertainty could be acceptable for empirical studies but not within a regulatory context where specific species' identification can be crucial, such as the identification of species at risk (Gilbey et al. 2021). Evaluation of false-positives in the detections of endangered or invasive species should include potential bias caused by gaps in reference libraries (Cristescu and Hebert 2018).

**Maximizing the reliability and the number of species assignments in eDNA metabarcoding studies by combining the use of a regional library and a public repository**

Our results showed that the use of a regional library increases both the reliability and number of species detected in an eDNA metabarcoding dataset. Yet, some species likely present in the Gulf of St. Lawrence were only detected with NCBI-nt, as discussed in the previous section. The growth of GSL-rl will increase the number of species that can be detected using the regional library, but unexpected species, such as new invasive species or species that have recently expanded their distribution, could remain undetected (Bohmann et al. 2014; Klymus et al. 2017; Piper et al. 2019; Stoeckle et al. 2020; Gold et al. 2021). Restricting species assignments to GSL-rl and avoiding the use of NCBI-nt would limit the maximum number of species detected.

Combining the strengths of a regional library with that of public repositories in a two-step approach is consequently the optimal solution to maximize reliability and number of species assigned in metabarcoding studies. Taxonomic assignments should be first performed with a regional library, ideally including a reliability ranking system as in the GSL-rl, to maximize the confidence in species assignment. We then strongly advise contrasting species assignment results from a regional library with those using a public repository to increase the number of species detections (see also Rohwer et al. 2018; Piper et al. 2019; Xiong et al. 2022 for similar recommendations). This would allow the reader to have a qualitative estimation of the species assignment accuracy. Species assignments relying uniquely on NCBI-nt should also clearly indicate that their reliability is limited.

We also encourage further benchmarking studies for the selection of optimal methods based on a broader comparison of assignment methods and the development of training sets for machine-learning methods. The choice between a more (e.g., BLAST-LCA) or less conservative approach (e.g., BLAST-TopHit) for species assignments should also reflect the study objectives. Our study had limited comparison of assignment methods. We selected methods often used in eDNA metabarcoding studies that are also performing relatively well in benchmarking assignment studies (O'Rourke et al. 2020; Hleap et al. 2021). We also compared assignments results between a curated regional library and NCBI-nt, which are opposed in their levels of curation. Performing similar analyses with other assignment methods (e.g., RDP, methods implemented in MEGAN CE, Huson et al. 2016) and using reference sequences resources with different levels of curation would be interesting. Our results also emphasize that future benchmarking studies should be done independently for regional libraries and public repositories, given the different properties of these resources, to maximize the reliability and the number of species assignments.

## Acknowledgements

We thank Grégoire Cortial and Jade Larivière for their inputs at the earlier stages of this study. We also thank Nick Jeffery, Christopher Hempel and two anonymous reviewers for helpful comments on previous versions of the manuscript. We thank Yanick Gendreau and Sandra Velasquez from the Coastal environmental baseline program and Geneviève Faille and Geneviève Côté from the Banc-des-Américains Marine Protected Area for eDNA sampling and the initial list of marine faunal species of interest.

# References


Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution 9(1): 134–147. https://doi.org/10.1111/2041-210X.12849

Aylagas E, Borja Á, Rodríguez-Ezpeleta N (2014) Environmental status assessment using DNA metabarcoding: Towards a genetics based marine biotic index (gAMBI). PLoS ONE 9(3): e90529. https://doi.org/10.1371/journal.pone.0090529

Bidartondo MI (2008) Preserving accuracy in GenBank. Science 319(5870): 1616. https://doi.org/10.1126/science.319.5870.1616a

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, de Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution 29(6): 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6(1): 1–17. https://doi.org/10.1186/s40168-018-0470-z

Bourdages H, Brassard C, Chamberland J-M, Desgagnés M, Galbraith P, Isabel L, Senay C (2022) DFO Can. Sci. Advis. Sec. Res. Doc. Preliminary results from the ecosystemic survey in August 2021 in the Estuary and northern Gulf of St. Lawrence. DFO.

Bush A, Compson ZG, Monk WA, Porter TM, Steeves R, Emilson E, Gagne N, Hajibabaei M, Roy M, Baird DJ (2019) Studying ecosystems with DNA metabarcoding: Lessons from biomonitoring of aquatic macroinvertebrates. Frontiers in Ecology and Evolution 7: 1–12. https://doi.org/10.3389/fevo.2019.00434

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13(7): 581–583. https://doi.org/10.1038/nmeth.3869

Camacho C, Madden T (2022) BLAST+ Release Notes. BLAST Help [Internet]. National Center for Biotechnology Information, Bethesda, MD. https://www.ncbi.nlm.nih.gov/books/NBK131777/

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10(1): 1–9. https://doi.org/10.1186/1471-2105-10-421

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Molecular Ecology 21(8): 1834–1847. https://doi.org/10.1111/j.1365-294X.2012.05550.x

Collins RA, Trauzzi G, Maltby KM, Gibson TI, Ratcliffe FC, Hallam J, Rainbird S, Maclaine J, Henderson PA, Sims DW, Mariani S, Genner MJ (2021) Meta-Fish-Lib: A generalised, dynamic DNA reference library pipeline for metabarcoding of fishes. Journal of Fish Biology 99(4): 1446–1454. https://doi.org/10.1111/jfb.14852

Costa FO, Landi M, Martins R, Costa MH, Costa ME, Carneiro M, Alves MJ, Steinke D, Carvalho GR (2012) A ranking system for reference libraries of DNA barcodes: Application to marine fish species from Portugal. PLoS ONE 7(4): 1–9. https://doi.org/10.1371/journal.pone.0035858

Cristescu ME, Hebert PDN (2018) Uses and misuses of environmental DNA in biodiversity science and conservation. Annual Review of Ecology, Evolution, and Systematics 49(1): 209–239. https://doi.org/10.1146/annurev-ecolsys-110617-062306

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology 26(21): 5872–5895. https://doi.org/10.1111/mec.14350

Delrieu-Trottin E, Williams JT, Pitassy D, Driskell A, Hubert N, Viviani J, Cribb TH, Espiau B, Galzin R, Kulbicki M, Lison de Loma T, Meyer C, Mourier J, Mou-Tham G, Parravicini V, Plantard P, Sasal P, Siu G, Tolou N, Veuille M, Weigt L, Planes S (2019) A DNA barcode reference library of French Polynesian shore fishes. Scientific Data 6(1): 1–8. https://doi.org/10.1038/s41597-019-0123-5

Edgar RC (2018) Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ 6: e4652. https://doi.org/10.7717/peerj.4652

Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO (2021) BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. Molecular Ecology Resources 21(2): 573–583. https://doi.org/10.1111/1755-0998.13262

Fraija-Fernández N, Bouquieaux MC, Rey A, Mendibil I, Cotano U, Irigoien X, Santos M, Rodríguez-Ezpeleta N (2020) Marine water environmental DNA metabarcoding provides a comprehensive fish diversity assessment and reveals spatial patterns in a large oceanic area. Ecology and Evolution 10(14): 7560–7584. https://doi.org/10.1002/ece3.6482

Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources 13(5): 851–861. https://doi.org/10.1111/1755-0998.12138

Gilbey J, Carvalho G, Castilho R, Coscia I, Coulson MW, Dahle G, Derycke S, Francisco SM, Helyar SJ, Johansen T, Junge C, Layton KKS, Martinsohn J, Matejusova I, Robalo JI, Rodríguez-Ezpeleta N, Silva G, Strammer I, Vasemägi A, Volckaert FAM (2021) Life in a drop: Sampling environmental DNA for marine fishery management and ecosystem monitoring. Marine Policy 124: 104331. https://doi.org/10.1016/j.marpol.2020.104331

Gold Z, Curd EE, Goodwin KD, Choi ES, Frable BW, Thompson AR, Walker Jr HJ, Burton RS, Kacev D, Martz LD, Barber PH (2021) Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. Molecular Ecology Resources 21(7): 2546–2564. https://doi.org/10.1111/1755-0998.13450

Hänfling B, Lawson Handley L, Read DS, Hahn C, Li J, Nichols P, Blackman RC, Oliver A, Winfield IJ (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Molecular Ecology 25(13): 3101–3119. https://doi.org/10.1111/mec.13660

Hering D, Borja A, Jones JI, Pont D, Boets P, Bouchez A, Bruce K, Drakare S, Hänfling B, Kahlert M, Leese F, Meissner K, Mergen P, Reyjol Y, Segurado P, Vogler A, Kelly M (2018) Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. Water Research 138: 192–205. https://doi.org/10.1016/j.watres.2018.03.003

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources 21(7): 2190–2203. https://doi.org/10.1111/1755-0998.13407

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R (2016) MEGAN Community Edition – Interactive exploration and analysis of large-scale microbiome sequencing data.

PLoS Computational Biology 12(6): 1–12. https://doi.org/10.1371/journal.pcbi.1004957

Jazdzewska AM, Tandberg AHS, Horton T, Brix S (2021) Global gap-analysis of amphipod barcode library. PeerJ 9: 1–28. https://doi.org/10.7717/peerj.12352

Kaehler BD, Bokulich NA, McDonald D, Knight R, Caporaso JG, Huttley GA (2019) Species abundance information improves sequence taxonomy classification accuracy. Nature Communications 10(1): 1–10. https://doi.org/10.1038/s41467-019-12669-6

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16(2): 111–120. https://doi.org/10.1007/BF01731581

Klymus KE, Marshall NT, Stepien CA (2017) Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. PLoS ONE 12(5): 1–24. https://doi.org/10.1371/journal.pone.0177643

Knebelsberger T, Landi M, Neumann H, Kloppmann M, Sell AF, Campbell PD, Laakmann S, Raupach MJ, Carvalho GR, Costa FO (2014) A reliable DNA barcode reference library for the identification of the North European shelf fish fauna. Molecular Ecology Resources 14: 1060–1071. https://doi.org/10.1111/1755-0998.12238

Layton KKS, Corstorphine EA, Hebert PDN (2016) Exploring canadian echinoderm diversity through DNA barcodes. PLoS ONE 11(11): 1–16. https://doi.org/10.1371/journal.pone.0166118

Leite BR, Vieira PE, Troncoso JS, Costa FO (2021) Comparing species detection success between molecular markers in DNA metabarcoding of coastal macroinvertebrates. Metabarcoding and Metagenomics 5: 249–260. https://doi.org/10.3897/mbmg.5.70063

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. Frontiers in Zoology 10(1): 1–14. https://doi.org/10.1186/1742-9994-10-34

Leray M, Knowlton N, Ho SL, Nguyen BN, Machida RJ (2019) GenBank is a reliable resource for 21st century biodiversity research. Proceedings of the National Academy of Sciences of the United States of America 116(45): 22651–22656. https://doi.org/10.1073/pnas.1911714116

Locatelli NS, McIntyre PB, Therkildsen NO, Baetscher DS (2020) GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. Proceedings of the National Academy of Sciences of the United States of America 117(51): 32211–32212. https://doi.org/10.1073/pnas.2007421117

Makiola A, Compson ZG, Baird DJ, Barnes MA, Boerlijst SP, Bouchez A, Brennan G, Bush A, Canard E, Cordier T, Creer S, Curry RA, David P, Dumbrell AJ, Gravel D, Hajibabaei M, Hayden B, van der Hoorn B, Jarne P, Jones JI, Karimi B, Keck F, Kelly M, Knot IE, Krol L, Massol F, Monk WA, Murphy J, Pawlowski J, Poisot T, Porter TM, Randall KC, Ransome E, Ravigné V, Raybould A, Robin S, Schrama M, Schatz B, Tamaddoni-Nezhad A, Trimbos KB, Vacher C, Vasselon V, Wood S, Woodward G, Bohan DA (2020) Key questions for next-generation biomonitoring. Frontiers in Environmental Science 7: 1–14. https://doi.org/10.3389/fenvs.2019.00197

Marques V, Milhau T, Albouy C, Dejean T, Manel S, Mouillot D, Juhel J-B (2021) GAPeDNA: Assessing and mapping global species gaps in genetic databases for metabarcoding studies. Diversity & Distributions 27(10): 1880–1892. https://doi.org/10.1111/ddi.13142

Mathon L, Valentini A, Guérin PE, Normandeau E, Noel C, Lionnet C, Boulanger E, Thuiller W, Bernatchez L, Mouillot D, Dejean T, Manel S (2021) Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. Molecular Ecology Resources 21(7): 2565–2579. https://doi.org/10.1111/1755-0998.13430

McCusker MR, Denti D, Guelpen L, Kenchington E, Bentzen P (2013) Barcoding Atlantic Canada's commonly encountered marine fishes. Molecular Ecology Resources 13(2): 177–188. https://doi.org/10.1111/1755-0998.12043

McGee KM, Robinson CV, Hajibabaei M (2019) Gaps in DNA-Based Biomonitoring Across the Globe. Frontiers in Ecology and Evolution 7: 1–7. https://doi.org/10.3389/fevo.2019.00337

Meiklejohn KA, Damaso N, Robertson JM (2019) Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. PLoS ONE 14(6): 1–14. https://doi.org/10.1371/journal.pone.0217084

Mioduchowska M, Czyż MJ, Gołdyn B, Kur J, Sell J (2018) Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too "universal"? PLoS ONE 13: e0199609. https://doi.org/10.1371/journal.pone.0199609

Murali A, Bhargava A, Wright ES (2018) IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. Microbiome 6(1): 140. https://doi.org/10.1186/s40168-018-0521-5

Nozères C (2017) Preliminary checklist of marine animal species of the Gulf of St. Lawrence, Canada, based on 4 sources. https://doi.org/10.13140/RG.2.2.10056.62727

O'Rourke DR, Bokulich NA, Jusino MA, MacManes MD, Foster JT (2020) A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. Ecology and Evolution 10(18): 9721–9739. https://doi.org/10.1002/ece3.6594

Oliveira LM, Knebelsberger T, Landi M, Soares P, Raupach MJ, Costa FO (2016) Assembling and auditing a comprehensive DNA barcode reference library for European marine fishes. Journal of Fish Biology 89(6): 2741–2754. https://doi.org/10.1111/jfb.13169

Pappalardo P, Collins AG, Pagenkopp Lohan KM, Hanson KM, Truskey SB, Jaeckle W, Ames CL, Goodheart JA, Bush SL, Biancani LM, Strong EE, Vecchione M, Harasewych MG, Reed K, Lin C, Hartil EC, Whelpley J, Blumberg J, Matterson K, Redmond NE, Becker A, Boyle MJ, Osborn KJ (2021) The role of taxonomic expertise in interpretation of metabarcoding studies. ICES Journal of Marine Science 78(9): 3397–3410. https://doi.org/10.1093/icesjms/fsab082

Paradis E, Schliep K (2019) Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics (Oxford, England) 35(3): 526–528. https://doi.org/10.1093/bioinformatics/bty633

Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, Blacket MJ (2019) Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience 8(8): 1–22. https://doi.org/10.1093/gigascience/giz092

Porter TM, Hajibabaei M (2018a) Automated high throughput animal CO1 metabarcode classification. Scientific Reports 8(1): 1–10. https://doi.org/10.1038/s41598-018-22505-4

Porter TM, Hajibabaei M (2018b) Over 2.5 million COI sequences in GenBank and growing. PLoS ONE 13: e0200177. https://doi.org/10.1371/journal.pone.0200177

Porter TM, Hajibabaei M (2020) Putting COI Metabarcoding in Context: The utility of exact sequence variants (ESVs) in biodiversity analysis. Frontiers in Ecology and Evolution 8: 1–15. https://doi.org/10.3389/fevo.2020.00248

Radulovici AE, Sainte-Marie B, Dufresne F (2009) DNA barcoding of marine crustaceans from the Estuary and Gulf of St Lawrence: A regional-scale approach. Molecular Ecology Resources 9: 181–187. https://doi.org/10.1111/j.1755-0998.2009.02643.x

Radulovici AE, Vieira PE, Duarte S, Teixeira MAL, Borges LMS, Deagle BE, Majaneva S, Redmond N, Schultz JA, Costa FO (2021) Revision and annotation of DNA barcode records for marine invertebrates: Report of the 8th iBOL conference hackathon. Metabarcoding and Metagenomics 5: 207–217. https://doi.org/10.3897/mbmg.5.67862

Ramirez JL, Rosas-Puchuri U, Cañedo RM, Alfaro-Shigueto J, Ayon P, Zelada-Mázmela E, Siccha-Ramirez R, Velez-Zuazo X (2020) DNA barcoding in the Southeast Pacific marine realm: Low coverage and geographic representation despite high diversity. PLoS ONE 15(12): 1–13. https://doi.org/10.1371/journal.pone.0244323

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System: Barcoding. Molecular Ecology Notes 7: 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. PLoS ONE 8(7): e66213. https://doi.org/10.1371/journal.pone.0066213

Richardson RT, Bengtsson-Palme J, Johnson RM (2017) Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Molecular Ecology Resources 17(4): 760–769. https://doi.org/10.1111/1755-0998.12628

Rohwer RR, Hamilton JJ, Newton RJ, McMahon KD (2018) TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. MSphere 3(5): e00327-18. https://doi.org/10.1128/mSphere.00327-18

Schenekar T, Schletterer M, Lecaudey LA, Weiss SJ (2020) Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. River Research and Applications 36(7): 1004–1013. https://doi.org/10.1002/rra.3610

Stoeckle MY, Soboleva L, Charlop-Powers Z (2017) Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. PLoS ONE 12(4): 1–15. https://doi.org/10.1371/journal.pone.0175186

Stoeckle MY, Das Mishu M, Charlop-Powers Z (2020) Improved environmental DNA reference library detects overlooked marine fishes in New Jersey, United States. Frontiers in Marine Science 7: 226. https://doi.org/10.3389/fmars.2020.00226

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21(8): 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Van Den Bulcke L, De Backer A, Ampe B, Maes S, Wittoeck J, Waegeman W, Hostens K, Derycke S (2021) Towards harmonization of DNA metabarcoding for monitoring marine macrobenthos: The effect of technical replicates and pooled DNA extractions on species detection. Metabarcoding and Metagenomics 5: 233–247. https://doi.org/10.3897/mbmg.5.71107

von Ammon U, Wood SA, Laroche O, Zaiko A, Tait L, Lavery S, Inglis GJ, Pochon X (2018) Combining morpho-taxonomy and metabarcoding enhances the detection of non-indigenous marine pests in biofouling communities. Scientific Reports 8(1): 1–11. https://doi.org/10.1038/s41598-018-34541-1

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73(16): 5261–5267. https://doi.org/10.1128/AEM.00062-07

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K-DB, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher JN, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. The Science of the Total Environment 678: 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

WoRMS Editorial Board (2020) World Register of Marine Species. https://doi.org/10.14284/170

Wright ES (2016) Using DECIPHER v2.0 to analyze big biological sequence data in R. The R Journal 8(1): 352–359. https://doi.org/10.32614/RJ-2016-025

Xiong F, Shu L, Zeng H, Gan X, He S, Peng Z (2022) Methodology for fish biodiversity monitoring with environmental DNA metabarcoding: The primers, databases and bioinformatic pipelines. Water Biology and Security 1(1): 100007. https://doi.org/10.1016/j.watbs.2022.100007

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3(4): 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zafeiropoulos H, Gargan L, Hintikka S, Pavloudi C, Carlsson J (2021) The Dark mAtteR iNvestigator (DARN) tool: Getting to know the known unknowns in COI amplicon data. Metabarcoding and Metagenomics 5: 163–174. https://doi.org/10.3897/mbmg.5.69657

## Supplementary material 1
## Creation of Gulf of St. Lawrence regional library (GSL-rl) and creation of an eDNA metabarcoding dataset

Authors: Audrey Bourret, Claude Nozères, Eric Parent, Geneviève J. Parent

Data type: 7z. Arhive

Explanation note: List of species retrieved in BOLD under different names. List of taxa BIN number removed. Steps to obtain, filter and select publicly available sequences. Metadata of the GSL-rl, including ranking systems. [csv file]. BINs shared by two or more taxa. Taxa sharing more than one BIN within the GSL-rl. Characteristic of the curated regional library of the Gulf of St. Lawrence (GSL-rl) version 1.0. Sampling locations for the metabarcoding dataset in the St. Lawrence. Schematic representation of the metabarcoding bioinformatics pipeline.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.7.98539.suppl1